Constructing pseudowords for experimental research: Problems and solutions*

Anthi Revithiadou, Dimitra Ioannou, Maria Chatzinikolaou & Katerina Aivazoglou

Aristotle University of Thessaloniki
revith@lit.auth.gr, dioannos@lit.auth.gr, cnmaria@lit.auth.gr, aaivazog@lit.auth.gr

Abstract
We present the methodology for the construction of pseudowords for an experiment that explores the Greek listeners' perception of nominal stress. Our goal was, first, to construct pseudonouns that sound native enough to the native speakers' ears and, second, to make the speakers' 'familiarity intuition' measurable. For this purpose, we created a noun-only version of the *Clean Corpus*/ILSP, named *NClean*. We also set a number of variables to evaluate the phonotactic familiarity of the constructed pseudonouns relative to the mean phonotactic characteristics of the NClean nouns. Pseudonouns that complied to the selected phonotactic criteria qualified as experimental items.

Keywords: Greek stress, phonotactics, pseudowords, experimental research, Clean Corpus

1. Setting the stage: Stress in Greek

Greek language has morphology-determined stress, that is, stress assignment is determined on the basis of grammar-specific principles and not on purely phonological ones. More specifically, the majority of morphemes, such as stems/roots, thematic vowels, derivational suffixes and inflectional endings have lexically-encoded accentual properties (i.e., they are accented, post-/pre-accenting, see Revithiadou 1999). Given that words in Greek consist of more than one morpheme, it is often the case that morphemes with different accentual properties may coexist in a word. Revithiadou (1999) has argued that when there is a conflict between accented morphemes, morphology offers a helping hand in deciding which accented morpheme will win. In the absence of lexically-encoded stress information,

* This research was supported by the 'Excellence 2011' Program [Action B, Ref.No 87883] awarded to Dr. Anthi Revithiadou and funded by the Research Committee of the Aristotle University of Thessaloniki.

however, stress is on the syllable dictated by the language-specific default, that is, on the antepenultimate, e.g., *píθikos* 'monkey', *krokóðilos* 'crocodile' (Malikouti-Drachman & Drachman 1989; Ralli & Touratzidis 1992; Revithiadou 1999, 2007; Burzio & Tantalou 2007, a.o.). Morphology-oriented stress, in combination with boundedness to the last three syllables of the word, yield only three possible stress patterns for Greek: antepenultimate (APU), penultimate (PU) and ultimate (U) stress (Malikouti-Drachman & Drachman 1989; Drachman & Malikouti-Drachman 1999):

(1a)	APU	píθikos	'monkey-NOM.SG'
(1b)	PU	tsobános	'shepherd- NOM.SG'
(1c)	U	maragós	'carpenter- NOM.SG'
(2a)	APU	ɣítonas	'neighbor-NOM.SG'
(2b)	PU	eónas	'century-NOM.SG'
(2c)	U	vasiljás	'king-NOM.SG'
(3a)	APU	ɣéfira	'bridge-NOM.SG'
(3b)	PU	elpíða	'hope-NOM.SG'
(3c)	U	aɣorá	'market-NOM.SG'

The examples in (4b-c) demonstrate lexically-inflicted stress patterns, whereas the example in (4a), in which an accentless morpheme is combined with an accentless inflectional ending, represents the *phonological default* (PDf).

(4a)	/ɣiton-as/	accentless root
(4b)	/eón-as/	accented root
(4c)	/vasilj^-as/	post-accenting root ('^' = non-local accent)

PDf is an analysis-specific construct, which means that a differentiation in the analysis may result to a different definition of the PDf within the same language. For example in Russian, two different patterns have been proposed to represent the language PDf: (a) default is word initial (Halle 1973, 1997; Kiparsky & Halle 1977; Melvold 1990) and (b) default is post-stem (Alderete 1999, 2001a, b).

Although APU has been acknowledged to represent the phonological aspect of the Greek stress system (Malikouti-Drachman & Drachman 1989; Ralli & Touratzidis 1992; Revithiadou 1999, 2007; Burzio & Tantalou 2007 among others), it has not

been experimentally shown that APU is the prevalent or statistically preferred pattern. On the contrary, Protopapas, Gerakaki and Alexandri (2006) argue that it is the least preferred stress pattern in reading tasks. In a similar vein, a number of recent studies have shown that APU stress is marginal in suffixless words, e.g. acronyms (see Nikolou, Revithiadou and Papadopoulou 2012; Topintzi & Kainada 2012), and in certain classes of inflected words, e.g. nouns in *-a* (Apostolouda 2012). In order to shed light on this issue, Revithiadou, Lengeris and Ioannou (2013) and Revithiadou & Lengeris (to appear) designed and carried out two perception experiments that aimed at exploring whether Greek speakers show a bias for a specific stress pattern (e.g., the PDf) and, if yes, whether this bias depends on morphological information. Our research questions focus exclusively on the stress behavior of nouns.

For the purposes of the experiment, we were required to construct 200 pseudowords from five major morphological classes: nouns ending in *-os, -o, -a, -as,* and *-i* (fem/neut), and of specific size and syllable structure. The pseudowords were constructed on the basis of actual words (5a). The segmental (C, V) positions that were subject to modification are underlined in (5b).

(5)	*actual words*		*target C/V positions*
a.	CCV.CV.CV	b.	<u>CCV.CV.C</u>V(C)
	CV.CV.CV		<u>CV.CV.C</u>V(C)
	CCV.CV		<u>CCV.C</u>V(C)
	CV.CV		<u>CV.C</u>V(C)

Our main concern for the data sets (i.e., pseudowords) used at the experiment was that they should be constructed in a way that they were not familiar but still sounded Greek enough to the Greek speakers' ears. For this purpose, we developed a specific methodology that exploits corpus-based tools that are freely available on the web, but at the same time takes into consideration the morphological classhood and the morphosyntactic category of words (i.e., nouns).

2. The methodology of pseudoword construction

2.1 Constructing a category-specific corpus

The construction of pseudowords for a production experiment on morphology-oriented stress was proven to be a quite challenging task. The major methodological

issue was to find a reliable way to estimate the degree of familiarity of the pseudowords, that is, to make the speakers' 'familiarity intuition' measurable. In order to achieve this goal, we relied on an existing corpus, namely *Clean*. The *Clean Corpus*, created by Protopapas and his colleagues, is a component of the "ILSP Psycho-Linguistic Resource" (http://speech.ilsp.gr/iplr, cf. Protopapas et al. 2010). It is a medium size corpus which contains 217.664 types (approximately 29.6 million tokens)[1] culled up mainly from newspapers, magazines, legal and literary texts, etc.

The major advantage of Clean is that it is freely accessible on the web and, more importantly, it comes with an on-line tool, the *NumTool* (http://speech.ilsp.gr/iplr/NumTool.aspx, see Protopapas et al. 2010), which provides quantitative measures for each letter string/word submitted to the system by the user. The variables that were relevant to our study are the following:

(6a) Bigram frequencies (phonemes only): i. Logmean bigram token frequency; ii. Logmean bigram type frequency.

(6b) Neighborhoods & cohorts: i. N phonological neighbors (replace only); ii. N phonological neighbors (replace, delete, insert, transpose); iii. Phonological Levenshtein distance 20.

Bigrams are pairs of adjacent items; in phonological representations bigrams refer to pairs of phonemes:[2]

> Bigram counts are calculated by first summing up all the occurrences (tokens) of each combination of two phones. Total bigram frequency is related to the difficulty with which an item can be read, as it reflects the familiarity of the reader with the combinations of phones exhibited by a given item (word) […].
>
> The neighbors of an item are items (words) of equal length that differ from the probe item by a single segment.
>
> (http://speech.ilsp.gr/iplr/documentation.htm)

[1] A *type* is the unique form of a word, while a *token* is any occurrence of that particular word.
[2] Detailed information on the nature and the calculation of the variables is available on the ILSP webpage (http://speech.ilsp.gr/iplr, see also Protopapas et al. 2010).

There were several variables concerning bigrams. We excluded variables that compute the orthographical representation of the word and concentrated only on those that evaluate inputs on the basis of their phonological representation. More specifically, we chose *Logmean bigram token frequency* and *Logmean bigram type frequency*, which focus solely on phonemes of tokens and types, respectively.

The variables in (6b i-ii) count the number of the phonological neighbors if one applies just replacement or replacement, deletion, insertion and transposition, respectively. The variable in (6b iii) is a less strict measure of phonological distance that calculates the mean phonological distance of the N (typically 20) nearest items.

The variables in (6) allow us to control whether the constructed words are close to but yet not too distant from existing ones. This is because Clean comes with a package of online tools (e.g., the NumTool) and downloadable material (in the form of .txt or .xls files) which contains a full processed word list with associated frequency of occurrence and detailed quantitative measures of all variables for each word of the corpus (http://speech.ilsp.gr/iplr/downloads.htm). This information allows us to assess how each constructed word fares phonotactically compared to the ones in Clean.

A major drawback of Clean, however, is that it does not provide any information on the morphological category (e.g., noun, verb, pronoun, etc.) of listed words, which is of absolute relevance to the study at hand, due to the morphology-oriented nature of Greek stress. As argued in Revithiadou (1999), there is a sharp difference in stress between verbs and nouns; for instance, nouns exhibit more accentual contrasts than verbs. Moreover, stress is transparently associated to morphological information in verbs but not in nouns. For example, past forms are almost exclusively associated with APU stress.[3]

The solution to this problem was to develop a finer-grained, noun-targeted version of Clean, named *NClean*. The new corpus consists of 13.324 (underived/non-compound) nouns, all culled up from Clean, version: ignoring stress. We relied on the stressless version of Clean because, given the aims of our study, we didn't wish the variables in (6) to take into consideration in their calculations information on the position of stress. The next step was to extract the valuable information contained in

[3] APU stress is affiliated with the PAST either because past inflections have been (traditionally) argued to require stress to surface on the APU syllable (Warburton 1970; Babiniotis 1972; Ralli 2005) or because a stressed proclitic or prefixal element is present in the past form (see van Oostendorp 2007, 2012 and Spyropoulos & Revithiadou 2009, 2011, respectively).

NClean and effectively exploit it in the construction of pseudowords for our experimental research.

2.2 Constructing the pseudowords

The 200 pseudowords created for our experimental tasks on noun stress were controlled for morphological classhood, size and syllable structure. More specifically, two- and three-syllable long nouns from the five noun classes: *-os, -o, -a, -as*, and *-i* (feminine and neuter) were constructed. We opted for simple syllable structures, namely, CV.CV(C), CV.CV.CV(C), CCV.CV(C), CCV.CV.CV(C), in order to avoid the possible interference of phonotactics in our experimental results. Tables 1 and 2 show examples of masculine and feminine nouns, respectively. The pseudowords were of course stressless since familiarity had to be assessed on the basis of the phonotactic make-up and, particularly, on the combinatorial configurations of the input strings.

-as	2σ	3σ
CV.CVC	θokas	
CCV.CVC	krefas	
CCV.CV.CVC		trivetas
CV.CV.CVC		lavenas

Table 1. Masculine pseudonouns in -as

-a	2σ	3σ
CV.CV	rova	
CCV.CV	spika	
CV.CV.CV		letoma
CCV.CV.CV		krixena

Table 2. Feminine pseudonouns in -a

The following procedure was followed for each constructed word:

- STEP 1: All data of the NClean Corpus nouns were categorized according to size and syllable structure. As a result, nouns of the same length and syllable structure were grouped together. Some representative examples are provided in (7) and (8).

(7) disyllabic CCV.CV nouns

	A	B	C	D	E	F	G
1	spel	phon	BGtokfreqPho	BGtypfreqPho	nNeiPho	nNeiRDITPho	PLD20
396	σκουφι	skufi	0,112	0,351	5	7	1.650
397	σκουφο	skufo	0,116	0,323	5	6	1.600
398	σκουφοι	skufi	0,112	0,351	5	7	1.650
399	σκουφου	skufu	0,068	0,200	4	5	1.700
400	σκυλα	scila	0,703	0,955	10	13	1.350
401	σκυλε	scile	0,711	0,904	7	7	1.600
402	σκυλι	scili	0,902	1,155	15	18	1.100
403	σκυλια	sciLa	0,147	0,241	8	8	1.650
404	σκυλιου	sciLu	0,083	0,133	6	6	1.750
405	σκυλο	scilo	0,745	1,026	13	16	1.150
406	σκυλοι	scili	0,902	1,155	15	18	1.100
407	σκυλου	scilu	0,516	0,688	10	12	1.400

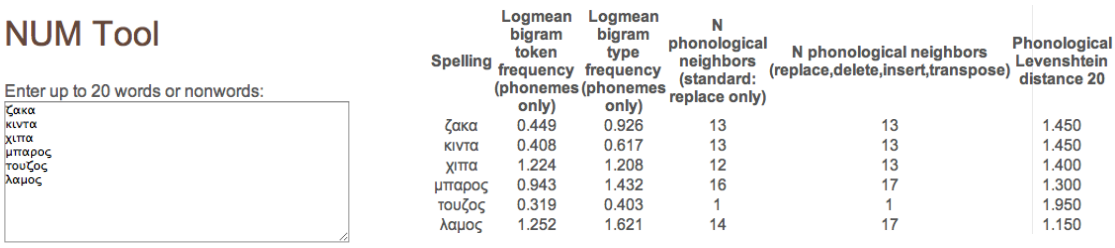
(8) trisyllabic CCV.CV.CVC nouns

	A	B	C	D	E	F	G
1	spel	phon	BGtokfreqPho	BGtypfreqPho	nNeiPho	nNeiRDITPho	PLD20
588	πριγκηπας	prihGipas	0,715	0,833	2	4	2150
589	πριγκηπες	prihGipes	0,703	0,778	2	2	2300
590	πριγκηπων	prihGipon	0,844	0,874	0	0	2600
591	πριγκιπας	prihGipas	0,715	0,833	2	4	2150
592	πριγκιπες	prihGipes	0,703	0,778	2	2	2300
593	πριγκιπων	prihGipon	0,844	0,874	0	0	2600
594	προβατον	provaton	1,511	1,616	0	1	1950
595	προβατων	provaton	1,511	1,616	0	1	1950
596	προβολεις	provolis	1,217	1,487	5	9	1550
597	προβολες	provoles	0,934	1,162	5	7	1650
598	προβολης	provolis	1,217	1,487	5	9	1550
599	προβολος	provolos	1,011	1,282	5	5	1750

- STEP 2: Mean values and SDs for bigram frequencies (phonemes only) and neighborhoods & cohorts were calculated anew for each noun category (e.g, for disyllabic CV.CVC nouns in *-os*, *-as*, disyllabic CCV.CV nouns in *-o*, *-a*, *-i*, etc., trisyllabic CCV.CV.CV nouns in *-o*, *-a*, *-i*, and so on). We set the acceptable range as strictly as possible from mean –1SD to mean +1SD. For instance, in nouns with syllable structure CV.CV, the mean value of BGtokfreqPho was 1,001 and the SD was 0,866. Thus, the permissible range was set from [mean value – SD = 1,001 - 0,866 =] **0,135** to [mean value + SD = 1,001 + 0,866 =] **1,867**. Similarly, the mean value of nNeiPho for the same category of nouns was 18 and the SD was 10. Hence the permissible range was set from [18 - 10 =] **8** to [8 + 10 =] **28**.
- STEP 3: Novel words were constructed and tested by the *NumTool*, which provided quantitative measures of the variables in question for each submitted word string.[4]

[4] The NumTool calculates the variables in (6) based on the Clean Corpus.

(9)



NUM Tool

Enter up to 20 words or nonwords:

ζακα
κιντα
χιπα
μπαρος
τουζος
λαμος

Spelling	Logmean bigram token frequency (phonemes only)	Logmean bigram type frequency (phonemes only)	N phonological neighbors (standard: replace only)	N phonological neighbors (replace,delete,insert,transpose)	Phonological Levenshtein distance 20
ζακα	0.449	0.926	13	13	1.450
κιντα	0.408	0.617	13	13	1.450
χιπα	1.224	1.208	12	13	1.400
μπαρος	0.943	1.432	16	17	1.300
τουζος	0.319	0.403	1	1	1.950
λαμος	1.252	1.621	14	17	1.150

- STEP 4: Words that fell within the defined range of *all* variables at hand (see Step 2) were selected as suitable items for the experiment. Those that failed to fit to the defined range of at least one variable were discarded as unsuitable.

Some representative examples of pseudowords are provided in Tables 3 and 4. Each table presents the range of each variable within the specific category (feminine *-a* and masculine *-os*, respectively). As mentioned above, in order for a constructed pseudoword to be accepted, its values for all variables at play should be within the appropriate range. Novel words whose values deviated from a given range (e.g., *τουζος* /tuzos/) were discarded as experimental items.

Feminine nouns in -a, 2σ, syllable type: CV.CV						
Pseudowords		BGtok freqPho	BGtyp freqPho	nNei Pho	nNei RDITPho	PLD20
		0,135-1,867	0,305-1,996	8-28	9-33	0,942-1,562
ζακα	zaka	0,449	0,926	13	13	1,450
κιντα	kida	0,408	0,617	13	13	1,450
χιπα	xipa	1,224	1,208	12	13	1,400

Table 3. Feminine pseudonouns in -a

Masculine nouns in -os, 2σ, syllable type: CV.CVC						
Pseudowords		BGtok freqPho	BGtyp freqPho	nNei Pho	nNei RDITPho	PLD20
		0,408-2,213	0,627-2,335	5-20	6-26	1,015-1,757
μπαρος	baros	0,943	1,432	16	17	1,300
~~τουζος~~	~~tuzos~~	0,319	0,403	1	1	1,950
λαμος	lamos	1,252	1,621	14	17	1,150

Table 4. Masculine pseudonouns in -os

The end result was a pool of words that complied to the defined value range of the respective word categories of the NClean Corpus. More importantly, their degree of familiarity to the native speakers' ears was well-defined and measurable, as intended.

3. Conclusions

This study demonstrates the usefulness of corpora and the associated quantitative tools in constructing experimental material that complies to the phonotactic restrictions and, in general, to the phonological structure of Greek. More importantly, it establishes a methodology that allows us to strictly define and compute in a principled way the degree of familiarity of novel words. Finally, it shows that enriching corpora with morphological information leads towards more targeted results and proves to be vital for experimental research that explores the nature of morphology-oriented stress.

References

Alderete, J. 1999. *Morphologically governed accent in optimality theory*. PhD dissertation, University of Massachusetts, Amherst.

Alderete, J. 2001a. *Morphologically governed accent in optimality theory*. New York: Routledge.

Alderete, J. 2001b. Dominance effects as transderivational anti-faithfulness. *Phonology* 18: 201-253.

Apostolouda, V. 2012. *Ο τόνος των ουσιαστικών της ελληνικής: Μια πειραματική προσέγγιση*. [*Nominal stress in Greek: An experimental approach*.] MA thesis, Aristotle University of Thessaloniki.

Babiniotis, G. 1972. *Το ρήμα της ελληνικής* [*The verb in Greek*]. Athens.

Burzio, L. & N. Tantalou. 2007. Modern Greek accent and faithfulness constraints in OT. *Lingua* 117: 1080-1124.

Drachman, G. & A. Malikouti-Drachman. 1999. Greek word accent. In H. van der Hulst (ed.), *Word prosodic systems in the languages of Europe*. Berlin & New York: Mouton de Gruyter, 897-945.

Halle, M. 1973. The accentuation of Russian words. *Language* 49: 312-348.

Halle, M. 1997. On stress and accent in Indo-European. *Language* 73: 275-313.

Kiparsky, P. & M. Halle. 1977. Toward a reconstruction of the Indo-European accent. In L.M. Hyman (ed.), *Studies in stress and accent*. Los Angeles: University of Southern California, 209-238.

Malikouti-Drachman, A. & G. Drachman. 1989. Stress in Greek. *Studies in Greek Linguistics 1989*. University of Thessaloniki, 127-143.

Melvold, J.L. 1990. *Structure and stress in the phonology of Russian*. PhD dissertation, MIT, Cambridge, MA.

Nikolou, K., A. Revithiadou & D. Papadopoulou. 2012. Exceptional stress patterns in the absence of morphological conditioning. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambaki-Vougioukli (eds), *10th International Conference on Greek Linguistics (ICGL10)*. Komitini: Democritus University of Thrace, 472-479.

van Oostendorp, M. 2007. Stress as a prefix in Modern Greek. Paper presented at *OCP 4*, January, 18-21, Rhodes.

van Oostendorp, M. 2012. Stress as a proclitic in Modern Greek. *Lingua* 122: 1165-1181.

Protopapas, A., S. Gerakaki & S. Alexandri. 2006. Lexical and default stress assignment in reading Greek. *Journal of Research in Reading* 29(4): 418-432.

Protopapas, A., M. Tzakosta, A. Chalamandaris & P. Tsiakoulis. 2010. IPLR: An online resource for Greek word-level and sublexical information. *Language Resources and Evaluation*. Retrieved 27 October 2013 from: http://link.springer.com/article/10.1007%2Fs10579-010-9130-z.

Ralli, A. 2005. *Μορφολογία* [*Morphology*]. Athens: Patakis.

Ralli, A. & L. Touratzidis. 1992. A computational treatment of stress in Greek inflected forms. *Language and Speech* 35: 435-453.

Revithiadou, A. 1999. *Headmost accent wins: Head dominance and ideal prosodic form in lexical accent systems.* PhD dissertation, LOT Dissertation Series 15 (HIL/Leiden University), The Hague: Holland Academic Graphics.

Revithiadou, A. 2007. Colored Turbid accents and containment: A case study from lexical stress. In S. Blaho, P. Bye & M. Krämer (eds), *Freedom of Analysis?*. Berlin and New York: Mouton de Gruyter, 149-174.

Revithiadou, A., A. Lengeris & D. Ioannou. 2013. In search of the default stress in Greek: Evidence from perception. Paper presented at *OCP10*, 16-19 January 2013, Boğaziçi University, Istanbul.

Revithiadou, A. & A. Lengeris. To appear. One or many? In search of the default stress in Greek. In J. Heinz, R. Goedemans & H. van der Hulst (eds), *Dimensions of linguistic stress.* Cambridge: Cambridge University Press.

Spyropoulos, V. & A. Revithiadou. 2009. The morphology of PAST in Greek. *Studies in Greek Linguistics* 29: 108-122.

Spyropoulos, V. & A. Revithiadou. 2011. *PAST in Greek: A case study of the interface between morphological structure and phonological realization*. Ms. University of Athens and University of Thessaloniki.

Topintzi, N. & E. Kainada. 2012. Acronyms and the placement of default stress in Greek. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambaki-Vougioukli (eds.), *10th International Conference on Greek Linguistics (ICGL10)*. Komitini: Democritus University of Thrace, 472-479.

Warburton, I. 1970. *On the verb of Modern Greek*. The Hague: Mouton and Co.